

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

**Uso de Cadeias de Markov para identificar
atribuições em Mídias Digitais**

Raquel Malheiro de Carvalho

Trabalho de Conclusão de Curso

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

Uso de Cadeias de Markov para identificar
atribuições em Mídias Digitais

Raquel Malheiro de Carvalho
Orientador: Márcio Luis Lanfredi Viola

Trabalho de Conclusão de Curso apresentado
como parte dos requisitos para obtenção do
título de Bacharel em Estatística.

São Carlos
Setembro de 2022

Raquel Malheiro de Carvalho

Uso de Cadeias de Markov para identificar
atribuições em Mídias Digitais

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Raquel Malheiro de Carvalho e aprovado pela banca examinadora.

Aprovado em 15 de setembro de 2022.

Banca Examinadora:

- Márcio Luis Lanfredi Viola (Orientador)
- Andressa Cerqueira
- Ricardo Felipe Ferreira

Resumo

Com o avanço do *Marketing* digital e o aumento de novas mídias pagas para a veiculação de campanhas e anúncios com a finalidade de vender produtos, muitos anunciantes se questionam sobre o uso de mídias que não são as finalizadoras da compra, ou seja, o usuário não converte nela. Desta forma, pode-se utilizar métodos estatísticos com a finalidade de identificar qual é a participação de uma mídia na conversão, considerando o comportamento dos usuários até que ocorra a conversão. Neste trabalho, utiliza-se Cadeias de Markov como um método de análise do comportamento de usuários em mídias digitais pela estimação das probabilidades de conversão dos usuários, para que seja possível, posteriormente, tomada de ações, como otimizações de campanhas, alocação de investimentos entre outras, para melhor entrega e retorno sobre o que se é investido.

Palavras-chave: *Cadeia de Markov, Marketing digital, Probabilidades de transição.*

Lista de Figuras

2.1	Exemplo de funil de vendas.	18
3.1	Mídias no caminho de conversão.	25
3.2	Caminhos de conversões do relatório do <i>Google Analytics</i> de caminho do agrupamento de canal.	25
3.3	Jornada do Usuário pelas Mídias no Caminho de conversão.	26
4.1	Diagrama de transição da primeira para segunda mídia no caminho de conversão.	31
4.2	Diagrama de transição a três passos.	33
4.3	Diagrama de transição a quatro passos.	34
4.4	Diagrama de transição a cinco passos.	35
4.5	Diagrama de transição a treze passos.	35

Lista de Tabelas

3.1	Jornadas de usuário e probabilidades de transição que podem ser interessantes em analisar.	28
4.1	Jornadas do usuário.	30
4.2	Médias na jornada do usuário e probabilidades de ocorrência.	31
4.3	Distribuição de frequências do número de passos até a conversão, de acordo com a jornada dos usuários.	32
4.4	Proporção em que a finalização das jornadas dos usuários ocorre em cada média.	36

Sumário

1	Introdução	13
1.1	Objetivo	15
1.2	Organização do Trabalho	15
2	Marketing Digital	17
2.1	Conceitos Básicos	17
2.2	Modelos de Atribuição	18
3	Cadeias de Markov	21
3.1	Processo Estocástico e Cadeias de Markov	21
3.2	Número de passos até a conversão	24
3.3	Atribuição baseada em cadeias de Markov de ordem 1	24
4	Resultados e Conclusões da Análise	29
4.1	Entendendo o banco de dados	29
4.2	Análise dos caminhos de conversão	30
4.3	Médias finais do caminho de conversão	36
5	Considerações Finais	37
	Referências Bibliográficas	39

Capítulo 1

Introdução

Atualmente, para quem divulga marca e produtos na *internet*, um problema comum é investir em mídias digitais sem que seja feita uma projeção devida e, também, sem saber qual caminho na *web* o usuário irá percorrer para encontrar seu produto e efetuar a compra.

Muitos profissionais de *marketing* digital, quando precisam realizar projeções de investimento das mídias, têm dificuldades em analisar os dados disponíveis. Em sua maioria, os dados dos sites são disponibilizados no *Google Analytics*.

Devido ao enorme número de relatórios com diversas métricas sobre localização, sessões e filtros, acabam se restringindo a analisar apenas dados históricos isolados, como custo, receita, taxa de conversão, custo por clique, que são indicadores relacionados a conversões, ou seja, vendas realizadas.

Diante desse cenário, muitas projeções são feitas sem que tenha sido aproveitado ao máximo a influência de uma mídia social no comportamento ou decisão de compra de um usuário em um produto de interesse. Esta é uma análise importante a se fazer dado que, cada vez mais, o consumidor está sendo influenciado pelas redes sociais no processo de decisão de compra ([Goulart, 2016](#)).

Conforme comentado, para a realização de uma análise deste comportamento, muitos profissionais acabam usando apenas os dados históricos, porém a projeção é realizada sem uma metodologia que explique a probabilidade de ocorrência de tal comportamento.

Dada a quantidade de modelos e utilizações, não é possível definir qual é o certo ou errado, mas é possível identificar qual o seu melhor uso, dependendo do objetivo da mídia, como: (i) *branding*, quando o objetivo é a apresentação da marca; (ii) consideração, quando o cliente já conhece a marca, mas pode considerar adquirir mais informações de

produtos; (iii) conversão, quando o cliente quer realizar a compra.

Otimizações, que são ações realizadas nas plataformas para modificar o orçamento de campanhas e outras métricas, para que o anúncio seja entregue nas melhores posições, ou seja, nas posições mais altas no *rank* do *Google*, ou nas demais mídias digitais, definem a importância da análise do modelo de atribuição em cada mídia e etapa do usuário. Isso tudo se dá com o intuito de, não apenas impactar usuários, mas obter receita, que é um dos maiores objetivos da etapa de conversão.

Além disso, há estratégias com mais de uma mídia, nas quais uma é utilizada para atrair usuários, ou seja, sem retorno financeiro, o chamado Retorno sobre Investimento (ROI), e a outra tem como foco o que é chamado fundo de funil, etapa de conversão em que o usuário efetua a compra e a empresa tem o retorno sobre o que foi investido.

A partir dessas análises, pode-se não somente identificar o caminho que o usuário faz e quais as mídias que participam deste caminho, mas também reavaliar as estratégias utilizadas em cada mídia aumentando, diminuindo ou parando de investir ou, ainda, mudando seu posicionamento e etapa do funil que está sendo utilizada.

Outra oportunidade associada à identificação do caminho de usuários e quais mídias participam com maior ou menor frequência, dependendo do comportamento deles, é a análise de elasticidade destes investimentos, ou seja, qual o ponto ótimo que se deve investir nessas mídias.

Com o avanço do *marketing* digital e aumento de novas mídias pagas para a veiculação de campanhas e anúncios com a finalidade de vender produtos, muitos anunciantes se questionam sobre o uso de mídias que não são as finalizadoras da compra, ou seja, o usuário não converte nela.

Desta forma, pode-se utilizar métodos estatísticos com a finalidade de identificar qual é a participação de uma mídia na conversão, considerando o comportamento dos usuários até que ocorra a conversão.

Este trabalho complementarará as explicações dadas no trabalho realizado sobre "Modelos de atribuição para mídias digitais com uso de Cadeias de Markov" (Breno, 2017), tentando expandir para outras mídias que não apenas o Google Ads.

1.1 Objetivo

O objetivo deste estudo é, a partir da utilização de Cadeias de Markov de ordem 1, inferir a participação das mídias pagas no caminho de usuários, identificando em qual etapa de um funil de marketing a(s) mídia(s) se encaixa(m) para conseguir direcionar melhor o(s) investimento(s) e trazer um retorno sobre o investimento maior.

Em outras palavras, o objetivo é obter qual é a probabilidade de um usuário realizar uma compra, ou seja, conversão em uma mídia, utilizando o histórico de cliques relevantes (caminhos relevantes). Desta forma, verificaremos quais mídias são relevantes para determinar o melhor caminho de conversão.

Espera-se que, a partir da análise de dados, utilizando Cadeias de Markov, profissionais de *marketing* digital possuam mais uma ferramenta para auxiliar no planejamento de mídias que não só *Google Analytics*.

1.2 Organização do Trabalho

Este trabalho está organizado em cinco capítulos. No Capítulo 2 serão introduzidos os conceitos mais importantes envolvendo *Marketing digital*, possibilitando a compreensão de alguns conceitos como atribuições, conversões e o que são os caminhos de conversão a serem analisados.

No Capítulo 3 será definida Cadeia de Markov, tendo como foco, as Cadeia de Markov de ordem 1. Ainda, no Capítulo 3, será comentado sobre a atribuição baseada em cadeias de Markov de ordem 1.

No Capítulo 4 está descrita a base de dados utilizada neste trabalho, além da sua análise. O Capítulo 5, último capítulo, apresenta as considerações finais relativas ao trabalho.

Capítulo 2

Marketing Digital

2.1 Conceitos Básicos

O conceito de *marketing digital* é muito subjetivo e engloba desde uma ação de uma empresa na *internet* até as melhores estratégias e como extrair o máximo potencial dessas ações.

Marketing digital é a promoção de produtos ou marcas por meio de mídias digitais e é uma das principais maneiras que as empresas têm para se comunicar com o público de forma direta, personalizada e no momento certo.

Porém, com as evoluções tecnológicas e mudanças no padrão de compra do consumidor, as formas para fazer com que isso aconteça, ou seja, de chegar ao consumidor, mudam quase que diariamente.

É por isso que o *marketing* digital é a principal forma de fazer e conceituar *marketing* nos dias de hoje. O termo é usado para resumir todos os esforços de *marketing* no ambiente *on-line*.

Usando canais digitais como *blogs*, *sites*, motores de busca, mídias sociais, *e-mail*, entre outros, as empresas tentam obter potenciais clientes e realizar os desejos e solucionar problemas de clientes.

No *marketing* digital, há inúmeras métricas para várias áreas que englobam a etapa de conversão. Esta etapa faz parte das estratégias de venda de produtos, que fica no que é chamado “fundo do funil de conversões”. O funil de conversões é ilustrado na Figura [2.1](#).

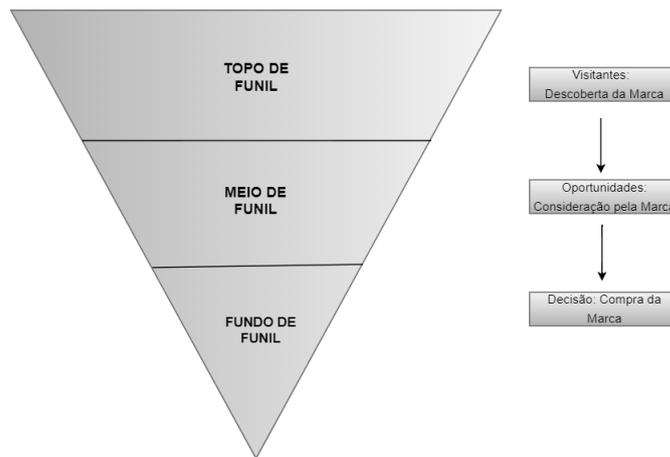


Figura 2.1: Exemplo de funil de vendas.

Tais etapas do funil abrem um leque de métricas a serem utilizadas para definir o desempenho, ou seja, desde a efetividade da entrega de um anúncio até a realização da compra feita pelo usuário. Dentre as métricas utilizadas, a que se destaca, devido a sua grande usabilidade no mercado de *e-commerce*, é o ROI (*Return over Investment*).

O ROI é utilizado para validar se o que é investido em publicidade está trazendo um retorno desejado, uma receita esperada, uma boa rentabilidade para o negócio. Com isso, outra métrica importante para esta análise é o investimento.

Para definir qual é o investimento “ideal” para uma campanha e, conseqüentemente, para a etapa do funil, são necessárias projeções por meio de planos de mídias, em que é levado em consideração o histórico presente nas plataformas de anúncios, bem como resultados obtidos e experiências passadas.

Devido a este processo, há oportunidade de analisar o comportamento das mídias e suas métricas de modo a entender quais os comportamentos possíveis que o usuário pode ter e como isto pode ser levado em consideração em qual mídia investir, investindo mais ou menos, para que se obtenha um bom resultado em receita, por exemplo.

2.2 Modelos de Atribuição

Para entender como são analisadas e como fluem os dados em uma das plataformas mais utilizadas em *marketing* digital, o *Google Analytics Universal* (GA3), é preciso compreender o que é e quais são os modelos de atribuição.

Modelo de atribuição é uma regra, ou conjunto de regras, utilizada para definir em

qual campanha e/ou mídia uma conversão deve ser atribuída.

Existem alguns tipos de modelos de atribuição dentro no *Google Analytics Universal*:

- Última interação;
- Último clique não direto;
- Último clique do Google Ads;
- Primeira interação;
- Modelo de atribuição Linear;
- Redução de Tempo;
- Baseado na Posição.

Sob os diferentes modelos de atribuição, o crédito é distribuído de maneira diferente e, assim, podem-se obter resultados diferentes a partir da tomada de decisão baseada em diferentes modelos.

Imagine a seguinte situação: com um modelo X , suas campanhas de pesquisa representam 85% das vendas, enquanto *display* e *YouTube* representam 15%. Já com o modelo Y , suas campanhas de pesquisa trazem 50% das vendas. Cada modelo de atribuição interpreta os mesmos resultados de maneiras diferentes a depender de sua estratégia.

Processos Markovianos podem ser utilizados como um modelo de atribuição e pode ser dito como o mais eficiente deles. É importante ressaltar que não há um modelo melhor que outro, porém um modelo usando cadeias de Markov possibilita estratégias mais completas na maioria dos casos.

A análise de Cadeias de Markov pode ser vista como um modelo de atribuição baseado em dados como o DDA (*Data Driven Attribution*), que nada mais é que a capacidade de processamento e interpretação massiva de dados aplicada à atribuição, que utiliza o Machine Learning do Google.

Este modelo atribui crédito a todos os canais envolvidos na jornada, ou seja, o caminho que do usuário, para uma conversão. Logo, uma atribuição usando cadeias de Markov permite que todas as diferentes mídias obtenham o crédito que merecem, por meio da análise de todos os caminhos de conversão, possibilitando entender o quanto cada um dos passos do usuário foi importante para gerar a conversão final.

Os modelos de atribuição mais comuns envolvem apenas um único ponto de contato, não considerando toda a jornada do consumidor e a quantidade de interações. Por exemplo, um modelo de atribuição de último clique atribuirá muita importância às mídias de fundo de funil e pouca importância para as mídias de assistência. Estes conceitos apresentados são baseados na página eletrônica do *Google*.

Para entender melhor como seria esta aplicação, precisamos entender o que são as Cadeias de Markov e como aplicá-las.

Capítulo 3

Cadeias de Markov

3.1 Processo Estocástico e Cadeias de Markov

Nesta seção são apresentadas as definições de processo estocástico e de um tipo especial de processo que corresponde aos Processos Markovianos ou Cadeias de Markov.

Um processo estocástico $\{X_t : t \in I\}$ (ou $(X_t)_{t \in I}$) é uma coleção de variáveis aleatórias definidas em um mesmo espaço de probabilidade $(\Omega, \mathcal{A}, \mathbf{P})$, indexadas por um conjunto de índices I , que, frequentemente, representa o tempo discreto. Neste caso, $I = \mathbf{N}$ ou $I = \mathbf{Z}$. Neste trabalho, será considerado $I = \mathbf{N}$.

Para cada tempo $t \in I$, X_t é uma variável aleatória, cujo conjunto formado pelos seus possíveis valores é chamado alfabeto e será denotado por \mathcal{B} .

Por exemplo, se a variável aleatória representa o valor da ação de uma determinada empresa e que o índice t indica o tempo, X_{t_1} é o valor da ação no tempo t_1 , X_{t_2} é o valor da ação no tempo t_2 , X_{t_3} é o valor da ação no tempo t_3 , e assim por diante. Note que, em cada índice t , a variável representa a mesma “característica” (valor da ação) e, na linguagem probabilística, isto é formalizado por “variáveis aleatórias definidas em um mesmo espaço mensurável de probabilidade”.

Em *marketing* digital, o processo estocástico pode representar o caminho que usuários percorrem até a conversão.

Cadeia de Markov é um processo estocástico vastamente estudado dada a sua aplicabilidade em um grande número de fenômenos físicos, biológicos, econômicos, sociais.

De modo geral, em uma Cadeia de Markov, a probabilidade de X_t depende dos valores das variáveis aleatórias $X_{t-1}, X_{t-2}, \dots, X_{t-k}$. Neste sentido, a quantidade k refere-se à estrutura (alcance) da dependência em relação aos valores passados para predizer o valor

no instante seguinte. Neste caso, k é a ordem do processo de Markov e, em geral, é finita.

Por exemplo, em uma cadeia de Markov de ordem um, a previsão do próximo passo, conhecendo toda a história passada do processo desde o seu início é tão boa quanto à previsão feita conhecendo-se apenas o valor do processo no presente.

Em uma cadeia de Markov de ordem $k \geq 1$, a previsão do próximo passo, conhecendo toda a história passada do processo desde o seu início é tão boa quanto à previsão feita conhecendo-se apenas os k valores do passado mais recente.

Especificada a ordem da cadeia, digamos $k \geq 1$, as sequências formadas por todos os possíveis valores que as variáveis $X_{t-1}, X_{t-2}, \dots, X_{t-k}$ possam assumir são chamados estados. O conjunto constituído pelas referidas sequências é denominado espaço de estados. Matematicamente,

$$\mathcal{S} = \underbrace{\mathcal{B} \times \dots \times \mathcal{B}}_{k \text{ vezes}} = \mathcal{B}^k.$$

A seguir, será definido cadeia de Markov de ordem 1 (Durrett, 1999; Ross, 1996).

Uma cadeia de Markov de ordem 1, $\{X_t : t \in \mathbf{N}\}$, satisfaz

$$P(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, X_{t-2} = i_{t-2}, X_0 = i_0) = P(X_{t+1} = j | X_t = i), \quad (3.1)$$

para todo $j, i, i_{t-1}, \dots, i_0 \in \mathcal{B}$ e para todo $t \in \mathbf{N}$.

Neste trabalho serão utilizadas cadeias de Markov homogêneas no tempo, ou seja, as probabilidades envolvidas em (3.1) são invariantes em relação a t . Assim, $\dots = P(X_{t+2} = j | X_{t+1} = i) = P(X_{t+1} = j | X_t = i) = P(X_t = j | X_{t-1} = i) = \dots = P(X_1 = j | X_0 = i)$, para todo $j, i \in \mathcal{B}$.

Em um processo de Markov de ordem 1, o espaço de estados \mathcal{S} coincide com o alfabeto, isto é, $\mathcal{S} = \mathcal{B}$.

As probabilidades $P(X_{t+1} = j | X_t = i)$, $j, i \in \mathcal{B}$, são denominadas probabilidades de transição e denotadas por P_{ij} . Elas dizem respeito à probabilidade de transição do estado i para o estado j em apenas um passo da cadeia.

A matriz $P = [P_{ij}]_{i,j \in \mathcal{S}}$ é chamada matriz de transição e satisfaz a condição $\sum_j P_{ij} = 1$, para cada $i \in \mathcal{S}$.

A probabilidade $P_{ij}^{(n)} = P(X_{t+n} = j | X_t = i)$ é a probabilidade de transição do estado i para o estado j em n passos da cadeia, ou seja, é a probabilidade do processo estar no

estado j depois de n passos da cadeia, dado que o processo estava no estado i .

A probabilidade $P_{ij}^{(n)}$ é calculada pelas equações de Chapman-Kolmogorov dadas por

$$P_{ij}^{(n+m)} = \sum_{k \in \mathcal{S}} P_{ik}^{(n)} P_{kj}^{(m)}.$$

No contexto de *marketing* digital, considerando a cadeia homogênea, as probabilidades de transição caracterizaram o comportamento do usuário até a conversão por meio das probabilidades de transição, ou seja, a probabilidade de um usuário realizar uma compra em uma mídia, dado o seu histórico relevante de cliques em outras mídias ou na mesma mídia em que houve a conversão. Desta forma, será avaliado qual é o caminho relevante do usuário associado a maior probabilidade de conversão.

Por exemplo, tendo dois caminhos possíveis de usuários, $M_1 \rightarrow M_2$ e $M_1 \rightarrow M_2 \rightarrow M_2$, as probabilidades referentes aos primeiro e segundo caminhos são, respectivamente,

$$P(X_1 = 2 | X_0 = 1)$$

e

$$P(X_2 = 2 | X_0 = 1, X_1 = 2).$$

Supondo um processo Markoviano, $P(X_2 = 2 | X_0 = 1, X_1 = 2) = P(X_2 = 2 | X_1 = 2)$. Se a probabilidade de transição dos caminhos $M_1 \rightarrow M_1$, $M_1 \rightarrow M_2$, $M_2 \rightarrow M_1$ e $M_2 \rightarrow M_2$ são 0,6, 0,4, 0,7 e 0,3 então

$$P = \begin{pmatrix} 0,6 & 0,4 \\ 0,7 & 0,3 \end{pmatrix}$$

Agora, será apresentada a generalização da definição de uma cadeia de Markov de ordem 1 para ordens maiores que um. Uma cadeia de Markov de ordem k , $k \geq 1$, satisfaz

$$\begin{aligned} P(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, X_{t-2} = i_{t-2}, X_0 = i_0) = \\ P(X_{t+1} = j | X_t = i, \dots, X_{t-k+1} = i_{t-k+1}), \end{aligned}$$

para todo $j, i, i_{t-1}, \dots, i_0 \in \mathcal{B}$ e para todo $t \in \mathbf{Z}$.

Conforme já mencionado, o espaço de estados associado a um processo de Markov de

ordem k é $\mathcal{S} = \mathcal{B}^k$.

3.2 Número de passos até a conversão

Dada uma jornada do consumidor, ou seja, o caminho percorrido até a conversão, o número de passos até a conversão varia de indivíduo para indivíduo e do momento que está sendo realizada a pesquisa. Assim, o número de passos até a conversão (tempo de conversão) é uma variável aleatória, aqui denotada por T .

Por exemplo, para o seguinte caminho de conversão realizado pelo usuário $M_1 \rightarrow M_2$ (observação da cadeia), o número de passos até a conversão (tempo de conversão) é 1, dado que este caminho foi realizado em 1 passo.

Desta forma, considerando o exemplo citado no parágrafo anterior, a observação M_1M_2 corresponde ao caminho $M_1 \rightarrow M_2$, sendo que a conversão ocorre na mídia terminal M_2 .

Neste trabalho, as observações do processo correspondem à sequência de mídias que o usuário acessa e a amostra é formada pelo caminho percorrido até a conversão de usuários. Vale ressaltar que serão consideradas apenas observações em que ocorra a conversão, a qual está associada à mídia terminal na sucessão de mídias que o usuário acessa.

3.3 Atribuição baseada em cadeias de Markov de ordem 1

Para que o leitor compreenda os conceitos até aqui expostos, é necessária a compreensão dos modelos de atribuição e, principalmente, como é o modelo baseado em cadeias de Markov.

Para fins ilustrativos, considere quatro mídias (Mídia 1, Mídia 2, Mídia 3 e Mídia 4) e a variável aleatória X_t , representando a mídia acessada pelo usuário no tempo t . Assim, X_t pode assumir um dos valores do alfabeto $\{1, 2, 3, 4\}$, em que 1, 2, 3 e 4 denotam, respectivamente, as Mídias 1, 2, 3 e 4.

Se a cadeia é de ordem 1, o espaço de estados é dado por $S = \{1, 2, 3, 4\}$.

Para uma cadeia de Markov de ordem 1, a probabilidade de transição a um passo representa a probabilidade do usuário acessar uma determinada mídia dado o conhecimento da mídia que ele acessou no tempo imediatamente anterior, ou seja, $P(X_1 = i | X_0 = j)$, $i, j \in \{1, 2, 3, 4\}$.

A Figura 3.1 mostra o diagrama de fluxo, indicando as transições permitidas de uma mídia a outra. Cada aresta está associada a uma probabilidade de transição.

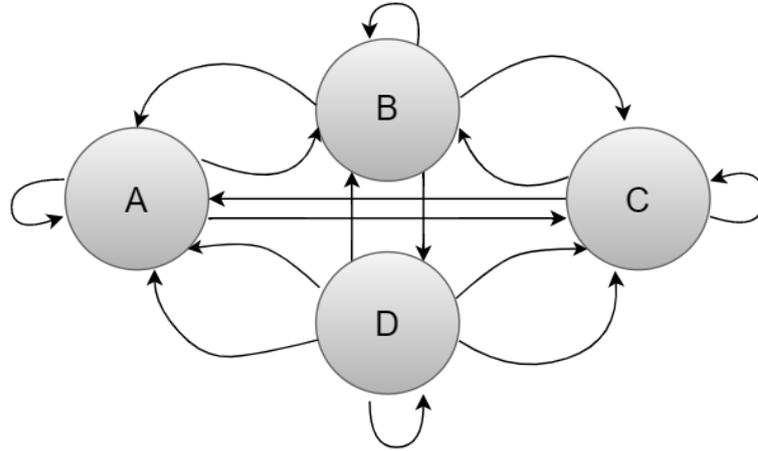


Figura 3.1: Mídias no caminho de conversão.

Antes do ocorrer a conversão, há diversos caminhos pelos quais o usuário pode percorrer de forma que uma mesma mídia pode ser acessada várias vezes. Isto será ilustrado por meio alguns exemplos.

Primeiramente, considere a Figura 3.2, a qual mostra o caminho que o usuário percorreu.

Caminho do Agrupamento de Canais do Funil Multicanal ?		Conversões ?
1.	Direto × 2	18 (15,93%)
2.	Direto × 3	9 (7,96%)
3.	Pesquisa Orgânica → Direto	9 (7,96%)
4.	Direto × 4	8 (7,08%)
5.	Direto × 5	6 (5,31%)
6.	Pesquisa Orgânica → Direto × 4	5 (4,42%)
7.	Direto × 6	4 (3,54%)
8.	Pesquisa Paga → Direto	4 (3,54%)
9.	Direto × 9	3 (2,65%)
10.	Pesquisa Paga → Pesquisa Orgânica	3 (2,65%)

Figura 3.2: Caminhos de conversões do relatório do *Google Analytics* de caminho do agrupamento de canal.

Estes caminhos são encontrados e podem ser extraídos no *Google Analytics* onde se encontram os relatórios de “Conversões”, sendo necessário acessar a parte de “Funis multi-canal” e, na sequência, “Principais caminhos de conversão”. Neste relatório, o “Caminho

do Agrupamento de Canais do Funil Multicanal” se refere aos caminhos, jornadas, que os usuários percorreram, sendo enumerado pela ordenação, neste caso, de “Conversões”, que são número de vezes que ocorreu aquela jornada e houve conversão. Analisando a (“Jornada 4”), o usuário começou pelo “Direto” e repetiu este caminho 4 vezes seguidas, ou seja, ele retornou 4 vezes ao caminho que realizou a primeira vez.

Analisando a coluna “Conversões” na Figura 3.2, observa-se que, para o caminho em que o usuário percorreu “Direto” quatro vezes (Jornada 4), houve oito conversões, ou seja, dentre as conversões totais, esta equivale a 7,08% das ocorridas. Sendo este “Direto” uma das possibilidades de “Mídia” que o usuário pode interagir para converter.

Agora, considere a Figura 3.3, que ilustra quatro possíveis caminhos até a conversão, isto é, mostra a sequência de mídias que os usuários utilizaram até a conversão, que ocorre na última mídia.



Figura 3.3: Jornada do Usuário pelas Mídias no Caminho de conversão.

O primeiro usuário (“Jornada 1”) interagiu com a Mídia 1 e houve a conversão no primeiro momento, enquanto que o segundo usuário (“Jornada 2”) interagiu, sucessivamente, com as Mídias 2 e 1, ocorrendo a conversão na Mídia 1.

No segundo caso, pode-se ter interesse na probabilidade

$$P(X_1 = 1 | X_0 = 2),$$

principalmente pelo fato de que a conversão ocorreu na Mídia 1 em um passo, a partir da Mídia 2.

No terceiro caso, o usuário iniciou a jornada na Mídia 2, a um passo foi para Mídia 3

e converteu, a dois passos, na Mídia 1. Assim, pode-se ter interesse na probabilidade

$$P(X_2 = 1|X_0 = 2, X_1 = 3)$$

No caso das Jornadas 2 e 4, o número de passos até a conversão (tempo de conversão) é 1 e 3, respectivamente.

Como estamos interessados no caminho até a conversão, será considerada a probabilidade de transição a t passos, levantando os valores mais frequentes de t , ou seja, os números de passos mais frequentes até a conversão, para analisar as jornadas.

Na prática, apenas conhecemos as transições permitidas de uma mídia a outra. Baseando-se em uma amostra de tamanho n , as probabilidades de transição são estimadas por meio da proporção amostral, ou seja,

$$\hat{P}(X_1 = i|X_0 = j) = \frac{N_n(ji)}{N_n(j)},$$

em que $N_n(ji)$ e $N_n(j)$ representam, respectivamente, o número de ocorrências das sequências ji e j na amostra de tamanho n e $i, j \in \mathcal{B}$.

A partir das probabilidades de transição estimadas, as probabilidades de transição a t passos serão calculadas, considerando os números de passos mais frequentes até a conversão. Com isto, teremos ideia dos caminhos mais prováveis.

A Tabela 3.1 mostra exemplos de possíveis caminhos que o usuário pode percorrer até a conversão (jornada). Para cada jornada, é indicada a probabilidade de transição que, possivelmente, será interessante em analisar.

Se, na amostra, a jornada $M_1 \rightarrow M_2$ for uma das mais frequentes, então será interessante calcular $P(X_1 = 2|X_0 = 1)$.

Caso a jornada $M_1 \rightarrow M_2 \rightarrow M_1 \rightarrow M_3 \rightarrow M_1$ for uma das mais frequentes, então será interessante calcular $P(X_4 = 1|X_0 = 1)$.

Tabela 3.1: Jornadas de usuário e probabilidades de transição que podem ser interessantes em analisar.

Jornada	Número de passos t	Probabilidade de Transição a t passos
$M_1 \rightarrow M_2$	1	$P(X_1 = 2 X_0 = 1)$
$M_1 \rightarrow M_2 \rightarrow M_1 \rightarrow M_3 \rightarrow M_1$	4	$P(X_4 = 1 X_0 = 1)$
$M_1 \rightarrow M_2 \rightarrow M_3 \rightarrow M_2$	3	$P(X_3 = 2 X_0 = 1)$
\vdots	\vdots	\vdots
$M_i \rightarrow \dots \rightarrow M_j$	t	$P(X_t = j X_0 = i)$

Capítulo 4

Resultados e Conclusões da Análise

Neste capítulo é mostrada a análise de um banco de dados real de uma empresa que investe em mídias pagas, tendo um histórico de dados de mais de 6 meses, supondo uma cadeia de Markov de ordem 1. A partir deste conjunto de dados, é feita a estimação das probabilidades de transição e da probabilidade de cada mídia, além de obter a distribuição do número de passos até a conversão.

Assim, a análise é feita considerando duas ações independentes: obtenção das probabilidades de transição estimadas e do número de passos mais frequentes até a conversão.

A partir das probabilidades de transição estimadas da Cadeia de Markov, será estudado o comportamento de interação do usuário na mídia e, assim, ter ideia das sequências de interações mais prováveis que levam à conversão, lembrando que a conversão representa quando o usuário realiza a compra.

4.1 Entendendo o banco de dados

O banco de dados foi obtido a partir dos relatórios do *Google Analytics*, em que é possível analisar no período de 6 meses. As informações obtidas correspondem a:

- Caminho de origem/mídia: as possíveis mídias pelas quais o usuário “passou”, o que chamamos neste estudo de “jornada”;
- Conversões: no caso de um *e-commerce*, as conversões são o número de compras efetuadas, ou outros tipo de ações definidas como conversões para a estratégia de negócio do anunciante como, por exemplo, número de adições ao carrinho, favoritar produtos, entre muitos outros;

- Valor das conversões

As mídias consideradas foram denotadas por A , B , C , D e E e o conjunto de dados é composto pela enumeração dos caminhos de conversão, ou seja, as jornadas. Por exemplo, $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ é uma das possíveis jornadas. Assim, os dados foram organizados de forma que cada observação corresponde à sequência de mídias que um usuário percorre até a conversão.

No caminho $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$, o usuário iniciou na mídia A , transitando, sucessivamente, pelas mídias B , C , D , até ocorrer a conversão na mídia E . Este caminho contribui na estimação da probabilidade das transições $A \rightarrow B$, $B \rightarrow C$, $C \rightarrow D$ e $D \rightarrow E$. Por outro lado, o caminho $A \rightarrow A \rightarrow A \rightarrow A \rightarrow A$ contribui apenas na estimação da probabilidade da transição $A \rightarrow A$.

A Tabela 4.1 mostra a forma de disposição dos dados.

Tabela 4.1: Jornadas do usuário.

Usuário	Jornadas
1	A E
2	B E
3	C A
...	...
32	A B B
33	B E E
34	A E E
...	...
80	B B B D
81	A A E E
...	...
706	AAAAAAAAAAAAAAAA

Foram selecionados apenas caminhos em que houve a conversão até 13 passos, totalizando 706 caminhos de conversão (jornadas).

4.2 Análise dos caminhos de conversão

Inicialmente, precisamos entender quantas vezes cada mídia participa dos caminhos de conversão, ou seja, a probabilidade estimada de ter a presença de uma das mídias nos dados, que é apresentada na Tabela 4.2.

Tabela 4.2: Mídias na jornada do usuário e probabilidades de ocorrência.

Mídia	Proporção
<i>A</i>	0,506
<i>B</i>	0,276
<i>C</i>	0,047
<i>D</i>	0,056
<i>E</i>	0,115

Nota-se, a partir da Tabela 4.2, que a mídia *A* aparece em, aproximadamente 51% das vezes no caminho de conversão dos usuários, seguida pela mídia *B* que representa cerca de 28% de chance de um usuário acessar tal mídia em um caminho de conversão. A mídia *E* aparece em última posição representando, aproximadamente, 12% de chance de ser acessada por um usuário.

Utilizando o pacote *Markovchain*, disponível no *Software R Studio*, foi possível obter a seguinte matriz de transição e o seu respectivo diagrama de transição, mostrado na Figura 4.1:

$$\begin{array}{c}
 \begin{array}{ccccc}
 & A & B & C & D & E \\
 A & \left[\begin{array}{ccccc}
 0,596 & 0,237 & 0,039 & 0,053 & 0,075 \\
 0,410 & 0,390 & 0,015 & 0,099 & 0,089 \\
 0,450 & 0,171 & 0,310 & 0,031 & 0,038 \\
 0,585 & 0,305 & 0,018 & 0 & 0,092 \\
 0,290 & 0,113 & 0,004 & 0,038 & 0,555
 \end{array} \right]
 \end{array}
 \end{array}$$

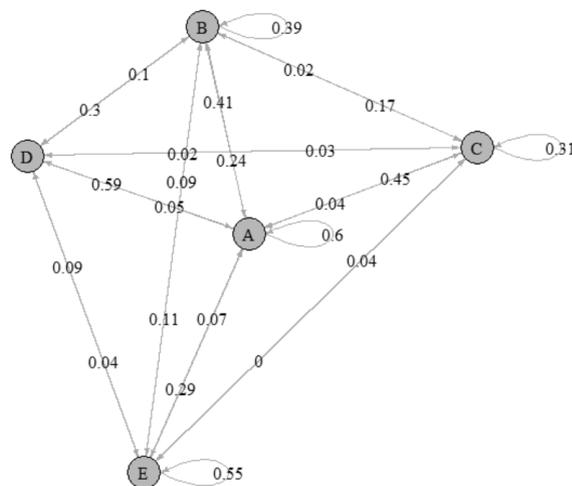


Figura 4.1: Diagrama de transição da primeira para segunda mídia no caminho de conversão.

No diagrama, por exemplo, a probabilidade de transição da mídia A para C é aquela próxima do nó de partida A , enquanto que a probabilidade de transição da mídia C para A é aquela próxima do nó de partida C . Assim, a probabilidade do usuário partir da mídia A para C é de 0,04, enquanto que partindo da mídia C para A é de 0,45, ou seja a transição $A \rightarrow C$ é muito mais provável.

Observa-se, a partir da matriz de transição, que a cadeia não possui apenas uma classe de estados, dado que o estado D não se comunicam entre si, portanto, não são irredutíveis. Em outras palavras, nota-se que a probabilidade do usuário ir da mídia D para a mídia D é zero, não ocorrendo a um passo. Porém, observa-se que a probabilidade estimada de um usuário sair da mídia A e ir para A , $P(X_1 = A|X_0 = A)$, é 0,596, ou seja, em 59,6% dos casos, o usuário sai da mídia A e retorna nela no passo seguinte. As maiores probabilidades associadas, neste caso, são aquelas em que os estados de partidas são as mídias A e D . Isso fica visível a partir do seu diagrama de transição representado na Figura 4.1.

Agora, ao analisar a distribuição do número de passos até a conversão, identificou-se o tamanho dos caminhos dos usuários que mais apareceram na base, ou seja, o número de passos mais frequente para que seja possível analisar as probabilidades de transição a partir de tal informação. A Tabela 4.3 mostra a frequência do número de cada passo até a conversão.

Tabela 4.3: Distribuição de frequências do número de passos até a conversão, de acordo com a jornada dos usuários.

Número de Passos	Frequência
1	24
2	84
3	155
4	137
5	107
6	70
7	42
8	26
9	15
10	10
11	5
12	3
13	5
15	2

Observa-se que os números de passos mais frequentes até a conversão são 2, 3, 4, 5 e 6 passos. Na maioria das jornadas, o usuário leva mais do que 1 passo e menos que 8 passos para converter. A maior probabilidade é termos uma jornada com 3 passos até a conversão e a segunda maior probabilidade está associada à jornada com 4 passos até a conversão.

Desta forma, obteve-se a matriz de transição em relação aos três passos mais frequentes até a conversão, que são 3, 4 e 5. Assim, procura-se entender as probabilidades de transição a t passos, em que $t = 3, 4, 5$, que estão associadas aos caminhos com o número de passos considerados até a conversão.

A 3 passos temos a seguinte matriz de transição e o diagrama mostrado na Figura 4.2:

	A	B	C	D	E
A	0.504	0.265	0.038	0.059	0.132
B	0.497	0.269	0.034	0.060	0.138
C	0.506	0.259	0.061	0.058	0.116
D	0.501	0.265	0.036	0.059	0.137
E	0.452	0.225	0.026	0.053	0.242

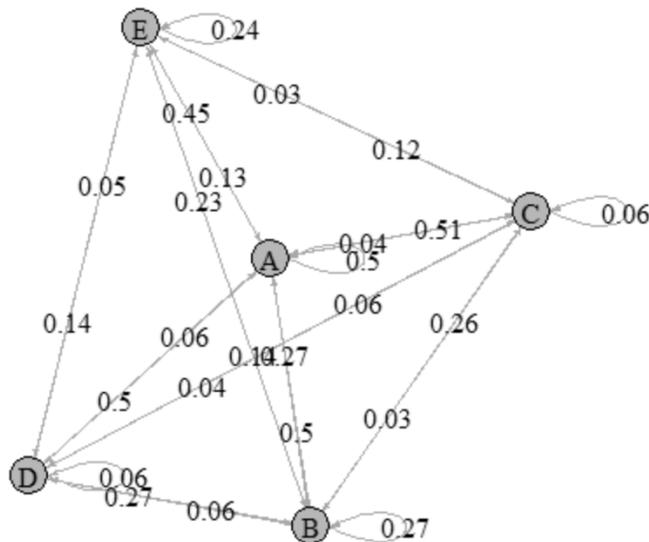


Figura 4.2: Diagrama de transição a três passos.

Observa-se que a cadeia possui apenas uma classe de estados, dado que, nesta, o estado D se comunica entre si, bem como os demais, e, portanto, são irredutíveis. Além disso, partindo do estado A , temos a probabilidade de 0,504 do usuário acessar o estado A depois de três passos. Como três passos até a conversão é o mais frequente, a maior

probabilidade é de que ocorra conversão na mídia A , partindo de qualquer outra mídia, em três passos.

A 4 passos temos a seguinte matriz de transição e o diagrama apresentado na Figura 4.3:

	A	B	C	D	E
A	0.4986050	0.2630349	0.03704226	0.05950424	0.1418137
B	0.4965952	0.2632110	0.03558892	0.05962040	0.1449845
C	0.5016274	0.2625471	0.04393809	0.05909490	0.1327926
D	0.4973781	0.2627137	0.03622381	0.05951730	0.1441670
E	0.4744069	0.2437854	0.03104381	0.05656786	0.1941961

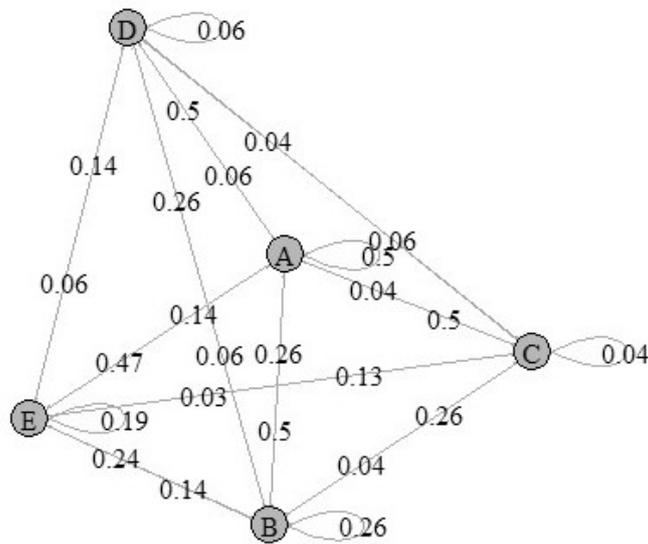


Figura 4.3: Diagrama de transição a quatro passos.

Observa-se que as transições mais prováveis a quatro passos são aquelas chegando a A , independentemente do estado de partida. Como quatro passos até a conversão é o segundo mais frequente, a segunda maior probabilidade é de que ocorra conversão na mídia A , partindo de qualquer outra mídia, em quatro passos.

A 5 passos temos a seguinte matriz de transição e o diagrama presente na Figura 4.4:

	A	B	C	D	E
A	0.4963738	0.2616041	0.03647266	0.05929753	0.1462520
B	0.4955808	0.2613438	0.03596252	0.05928220	0.1478306
C	0.4982235	0.2621576	0.03867477	0.05928371	0.1416605
D	0.4958343	0.2613190	0.03617667	0.05926359	0.1474064
E	0.4849225	0.2523645	0.03354854	0.05788761	0.1712769

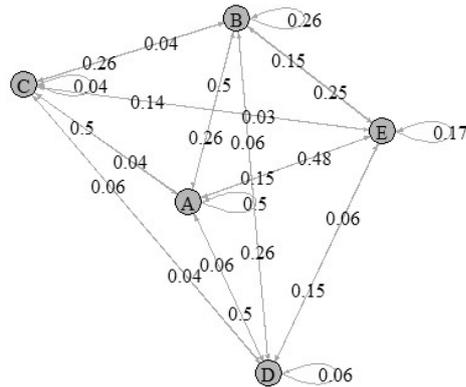


Figura 4.4: Diagrama de transição a cinco passos.

Note que as probabilidades de transição associadas a um mesmo estado de chegada são, praticamente, iguais, ou seja, são independentes da média de partida. Isto é uma evidência de que obteve-se a distribuição estacionária relativa às médias. Como cinco passos até a conversão é o terceiro mais frequente, a terceira maior probabilidade é de que ocorra conversão na média A , partindo de qualquer outra média, em cinco passos.

Dado o comportamento da matriz de transição a 5 passos, houve interesse em analisar o comportamento a 13 passos, cuja matriz de transição é apresentada a seguir e o diagrama é mostrado na Figura 4.5.

	A	B	C	D	E
A	0.4944858	0.2601545	0.03596341	0.05907971	0.1503166
B	0.4944837	0.2601529	0.03596285	0.05907946	0.1503211
C	0.4944919	0.2601595	0.03596511	0.05908047	0.1503030
D	0.4944843	0.2601534	0.03596301	0.05907953	0.1503198
E	0.4944548	0.2601294	0.03595527	0.05907589	0.1503846

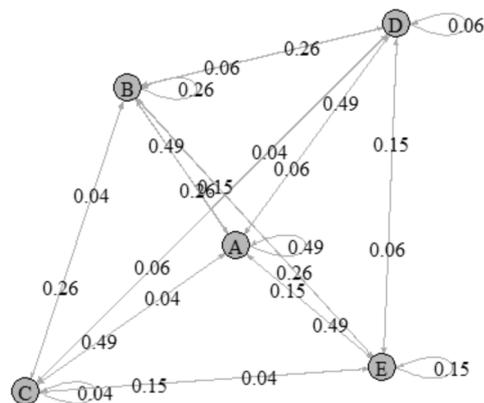


Figura 4.5: Diagrama de transição a treze passos.

Observa-se que, a um número maior de passos, as probabilidades de transição para uma média de chegada independe da média anterior, cujas probabilidades são, praticamente, iguais.

Quando o número de passos aumenta, as probabilidades associadas às colunas da matriz de transição vão se tornando iguais. Ou seja, as probabilidades de transição independem dos estados de partida.

4.3 Médias finais do caminho de conversão

A partir dos dados, obteve-se a proporção das médias finais. A Tabela 4.4 apresenta a proporção das médias em que ocorreu a conversão.

Tabela 4.4: Proporção em que a finalização das jornadas dos usuários ocorre em cada média.

Média	Finalizações	Proporção
A	340	0,48
B	178	0,25
C	31	0,04
D	45	0,06
E	112	0,16
Total	706	1,00

Nota-se que a média que mais obteve conversão, como era esperado de acordo com as análises anteriores, é a média *A*, sendo esta média a responsável por 48% das conversões, independente da média de partida. Já a média que representa o menor número de conversões é a média *C*, com 4% das conversões.

Observe que a proporção de cada média de conversão é muito próxima das probabilidades associadas à respectiva coluna na matriz de transição.

Capítulo 5

Considerações Finais

Este trabalho teve por objetivo aplicar cadeias de Markov de ordem 1 em um problema de *marketing* digital. Especificamente, o interesse foi inferir a participação das mídias pagas no caminho de usuários. Inicialmente, foram apresentados conceitos referentes à *marketing* digital e um exemplo de como podemos aplicar um processo Markoviano homogêneo de ordem 1 para modelar a sequência de mídias até a conversão.

Após isso, os conceitos de cadeias de Markov foram abordados, explicando os procedimentos a serem aplicados na análise e dando exemplos.

A partir da análise realizada em uma amostra de um banco de dados extraído do *Googla Analytics*, foi possível analisar a participação das mídias selecionadas, A, B, C, D e E , no caminho de conversão dos usuários.

Esta análise foi realizada estimando as probabilidades de transição e, em paralelo, obtendo a distribuição de frequências do número de passos mais frequentes até a conversão.

Os três passos mais frequentes até a conversão são 3, 4 e 5. Pelas probabilidades de transição estimadas, as maiores probabilidades correspondem à transição tendo a mídia A como chegada, independentemente da mídia de partida.

Na sequência, foram obtidas as probabilidades de transição a t passos, $t = 3, 4, 5$ (três passos mais frequentes até a conversão) e, adicionalmente, $t = 13$.

Por meio dos resultados obtidos, observou-se que a transição chegando a A , independentemente da mídia de partida, é a mais provável para o número de passos considerados. Além disso, como três passos até a conversão e o mais frequente, a maior probabilidade é de que ocorra conversão na mídia A , partindo de qualquer outra mídia, em três passos. Isto pode se dar pelo fato da mídia A estar presente no maior número de transições, participando no maior número de caminhos.

Além disso, a um número maior de passos, as probabilidades de transição para uma mídia de chegada independe da mídia anterior, cujas probabilidades são, praticamente, iguais, evidenciando que a obtenção da distribuição estacionária das mídias.

Por fim, obteve-se a proporção das mídias finais. A mídia que mais obteve conversão é a mídia *A*, sendo esta mídia a responsável por 48% das conversões, independente da mídia de partida.

Partindo de uma ótica de um profissional de *marketing* digital, com os resultados obtidos a partir da análise, é possível verificar quais são estas mídias, se os investimentos nelas estão sendo realizados de forma eficiente para alcançar as metas do seu negócio e, assim, otimizá-lo de acordo com as estimativas encontradas e objetivo da mídia.

Referências Bibliográficas

Breno, R. S. N. (2017). Modelos de atribuição para mídias digitais com uso de cadeias de markov: Um estudo de caso da agência de marketing digital. *Trabalho de Conclusão de Curso apresentado à Escola de Engenharia de São Carlos, da Universidade de São Paulo*, páginas 1–56.

Durrett, R. (1999). *Essentials of stochastic processes (Vol. 1)*. Springer.

Goulart, L. D. (2016). A influência das redes sociais no processo de decisão de compra. *V Seminário Científico do UNIFACIG*, **5**, 1–14.

Ross, S. M. (1996). *Stochastic processes*. John Wiley & Sons.